(14:06:32) David Gucwa: the thing about doing it from php is that it doesn't remember that you're logged in, so you can't go to the login page and then go to the search page
(14:06:45) David Gucwa: you have to go right to the search page and encode the login information into the url
(14:07:04) David Gucwa: thefacebook doesn't seem to like that for the search page for some reason
(14:08:27) dr ttol: whats a solution
(14:08:33) David Gucwa: i'm looking for one
(14:11:07) David Gucwa: you mentioned that when importing, you want to be able to send invite emails to people who aren't in connectu yet
(14:11:25) dr ttol: yes
(14:11:44) David Gucwa: I'm not sure how vital that is, but if we don't add that functionality then we will not need to crawl into other people's profiles and importing will take 5 seconds instead of 5 minutes
(14:12:18) David Gucwa: we can still get a list of their friends by id and find those friends on connectu if they have profiles already
(14:12:23) dr ttol: well
(14:12:25) dr ttol: if we crawl now
(14:12:31) dr ttol: then we'll also know who they're friends with
(14:13:01) David Gucwa: I think it would be a lot less trouble to do it at import time
(14:13:06) dr ttol: so if we have software that can crawl through every profile on a school, we'll know all the social networks
(14:14:04) dr ttol: there are two approaches to this
(14:14:09) dr ttol: we can either:
(14:14:20) dr ttol: a) index through "social net"
(14:14:37) dr ttol: b) index at import time, and import allt heir friends and all the profile information of themselves and their friends
(14:14:44) dr ttol: both will likely have the same outcome -- an index of the entire net
(14:14:51) David Gucwa: I'm not sure why we need the profile information for their friends
(14:14:53) David Gucwa: at import time
(14:15:54) dr ttol: because when that friend signs up, we already have their profile information
(14:16:53) David Gucwa: we can go fetch it pretty quickly when that friend signs up
(14:18:05) dr ttol: we still need the email addresses
(14:19:02) David Gucwa: That's what I was asking, how important is that functionality
(14:19:10) dr ttol: pretty important
(14:19:38) David Gucwa: which is more important, a short import time or sending email invites to non-users
(14:19:52) dr ttol: cant we send the email invites later
(14:20:06) dr ttol: it'll still be short import time, we can just have another software do the crawling of non-users
(14:20:18) dr ttol: we want to be able to send friend requests for email addresses already in our database
(14:20:20) David Gucwa: is the invite coming from connectu or from the person who just signed up
(14:20:30) dr ttol: connectu
(14:20:36) David Gucwa: okay that's fine then
(14:20:46) dr ttol: but via importer software
(14:20:53) dr ttol: like a sendmail()
(14:21:12) David Gucwa: right
(14:21:25) dr ttol: but we still need to crawl

Conversation with drttol at 2005-01-27 12:14:59 on David Gucwa (aim)
(12:14:59) dr ttol: hi
(12:17:21) David Gucwa: hi
(12:17:21) dr ttol <AUTO-REPLY>: We're upgrading our servers, more info: http://welcome.i2hub.com :)
(12:18:30) dr ttol: status update?
(12:19:21) David Gucwa: working on caching at the moment. email grabbing is functional.
(12:19:30) David Gucwa: Have you gotten a chance to test out the import page?
(12:19:32) dr ttol: ok, we have a better method
(12:19:35) dr ttol: to cache
(12:19:35) David Gucwa: ok
(12:19:47) dr ttol: if you log in
(12:19:50) dr ttol: then go to advanced search
(12:19:54) dr ttol: then search for all males
(12:19:56) dr ttol: then all females
(12:19:59) dr ttol: they are alphabeticized
(12:20:22) dr ttol: also, do you try to import relationship status as well?
(12:21:39) David Gucwa: I think so, I'm check that out
(12:21:45) David Gucwa: yeah I do
(12:22:30) dr ttol: cameron said it didnt do it correctly
(12:22:32) dr ttol: i mean
(12:22:42) dr ttol: do you try to see if that person is on connectu as well
(12:23:56) David Gucwa: does it tell you who the person is in a relationship with?
(12:24:13) dr ttol: yes
(12:24:18) dr ttol: tfb does
(12:25:36) David Gucwa: I can't find an example of that
(12:25:51) David Gucwa: oh okay
(12:25:59) David Gucwa: yeah sometimes it says
(12:28:17) David Gucwa: so do you want me to iterate through all of the search results to cache them all in advance?
(12:33:19) dr ttol: yeah, create a spider that does that
(12:33:28) dr ttol: start caching now, your other code will make use of the cache already right
(12:33:59) David Gucwa: yeah
(12:34:10) David Gucwa: it's difficult to retrieve their friends like this
(12:34:27) David Gucwa: it's very easy to get a list of friends for the profile that you log in with, because it displays them all on one page
(12:34:46) David Gucwa: for someone else's profile, it shows 10 to a page
(12:35:06) dr ttol: retrieve friends like what?
(12:35:38) David Gucwa: like it'll say "X has Y friends at Wellesley" with a link to see them
(12:35:56) dr ttol: whats the fastest way to do everything
(12:36:00) David Gucwa: I guess that's not entirely necessary to cache their friends
(12:36:11) David Gucwa: yeah I don't need to do that
(12:36:24) David Gucwa: we only need their friends when they're importing, then I can grab them from their own profile
(12:36:39) dr ttol: ok
(12:37:50) David Gucwa: just so I'm not going crazy, we all agreed that wellesley didn't display email addresses, didn't we?
(12:37:55) David Gucwa: because it's displaying now
(12:38:55) dr ttol: yeah
(12:38:56) dr ttol: odd
(12:39:00) dr ttol: thats why we want to cache
(12:39:04) dr ttol: the email addresses are KEY

Conversation with drttol at 2005-01-27 13:49:32 on David Gucwa (aim)
(13:49:32) dr ttol: status?
(13:50:21) David Gucwa: still working on the crawler
(14:39:03) David Gucwa: What do you think the maximum number of students in a given school is on the facebook?
(14:40:25) dr ttol: i dont know
(14:41:13) David Gucwa: I wrote a script that fetches a bunch of search results in parallel to make it a lot faster
(14:41:31) dr ttol: ok
(14:41:36) dr ttol: how fast can we get results now
(14:41:39) David Gucwa: but that means I don't know what the max number of search results is because I don't wait for a previous page to do the next one
(14:41:50) David Gucwa: so I just fetch up to the first 10k users
(14:41:54) David Gucwa: I figure that's enough
(14:42:07) dr ttol: ok
(14:42:08) dr ttol: if its not, what happens
(14:45:50) David Gucwa: i'm running speed tests at the moment
(14:46:01) dr ttol: ok
(14:46:05) dr ttol: what happens if there are 11k users
(14:46:07) David Gucwa: that huge delay between my IMs was due to all the cpu power on my computer getting sucked
(14:46:10) dr ttol: what happens to the missing 1k
(14:46:20) David Gucwa: then I don't index them
(14:46:27) dr ttol: and what happens then?
(14:46:32) dr ttol: they are indexed real time, if possible?
(14:46:49) David Gucwa: well I'd like to just get them all out of the way in advance
(14:47:02) David Gucwa: I could set a ridiculously huge upper bound to ensure that I get them all
(14:47:12) David Gucwa: I'm just wondering what a reasonable upper bound is
(14:47:22) dr ttol: we dont want to set an upper bound
(14:47:31) dr ttol: we want to seed connectu as much as possible
(14:47:41) dr ttol: but in the event that new users sign up on thefacebook after we've indexed
(14:47:46) dr ttol: we still want the code to go out and grab it
(14:47:51) dr ttol: (not the crawler, the imporer)
(14:47:57) dr ttol: importer
(14:48:03) David Gucwa: okay well the importer will still be doing some fetching then
(14:48:09) dr ttol: ok good
(14:48:14) David Gucwa: then it's not so bad if the crawler misses some
(14:48:18) dr ttol: the crawler is just to seed it so the importer can be faster too
(14:48:22) dr ttol: right
(14:50:43) David Gucwa: do we have a list of logins for each school that I can use
(14:51:17) dr ttol: we'll try to get you that
(14:51:24) dr ttol: right now, no, just try to index as much as possible
(14:51:30) dr ttol: with the list we gave you so far
(14:52:15) David Gucwa: ok
(14:54:19) David Gucwa: some good news is that I'm sending thefacebook several thousand requests in a couple minutes and they haven't blocked my IP yet
(14:55:31) dr ttol: ok
(14:58:16) dr ttol: right after importer is completed, we'll have you implement features to break others from doing the same to us
(14:58:28) David Gucwa: ok
(15:56:11) dr ttol: hi

(15:56:16) dr ttol: whats the url for stallscribbles rss
(15:57:09) David Gucwa: http://stallscribbles.com/rss/rss.xml
(16:16:19) David Gucwa: it just stopped showing emails again
(16:17:06) dr ttol: odd
(16:17:09) dr ttol: thats why we need to eindex
(16:17:10) dr ttol: asap
(16:17:15) dr ttol: did we get any
(16:17:20) David Gucwa: no
(16:17:26) David Gucwa: I think it stopped a little while ago
(16:17:36) dr ttol: ok, lets index other schools
(16:18:37) David Gucwa: k

Conversation with drttol at 2005-01-27 16:44:12 on David Gucwa (aim)
(16:44:12) dr ttol: status?
(16:44:23) David Gucwa: caching umich
(16:44:36) dr ttol: how fast?
(16:44:37) David Gucwa: I'm going to put the script on the server and see if it goes any faster
(16:44:43) David Gucwa: it's been going for like 10 minutes
(16:44:49) dr ttol: how many profiles
(16:45:11) dr ttol: tfb is run by a techie, so becareful
(16:45:25) David Gucwa: there are two steps, it retrieves all the id's first, then it gets the profile for each id
(16:45:30) David Gucwa: it's still on the first step, up to 6000 ids
(16:45:51) dr ttol: ok, does it get the pictures as well
(16:46:29) David Gucwa: it gets the urls of the pictures
(16:46:44) dr ttol: what about for the importer
(16:48:23) David Gucwa: same
(16:48:42) David Gucwa: it displays the pictures when you're choosing which one you want, but it doesn't store the picture anywhere locally yet
(16:49:02) dr ttol: when they pick the picture, does it store it locally
(16:49:05) dr ttol: to the cu database
(16:49:24) David Gucwa: no
(16:49:30) David Gucwa: I don't know where they're supposed to go
(16:49:37) dr ttol: ask joel
(16:50:04) David Gucwa: yeah that's something I'll need to take care of eventually
(16:50:13) David Gucwa: but I'm trying to get the crawler finished up for now
(16:50:27) dr ttol: ok
(17:27:05) dr ttol: status
(17:27:11) David Gucwa: importing 4 schools
(17:27:16) David Gucwa: it's taking a while
(17:27:17) dr ttol: all profiles?
(17:27:26) David Gucwa: I think it's actually going slower on the server
(17:27:28) David Gucwa: than on my computer
(17:27:35) David Gucwa: still just grabbing id's
(17:28:00) dr ttol: hm
(17:28:03) dr ttol: which server are you using
(17:28:19) David Gucwa: 69.44.59.182
(17:28:27) dr ttol: it could e that the schoools have more ids than you think
(17:28:40) dr ttol: so the crawler works?
(17:28:47) David Gucwa: yeah
(17:29:22) dr ttol: which universities
(17:29:24) dr ttol: do we have umass
(17:29:29) David Gucwa: no I don't have a umass login
(17:29:38) dr ttol: rdegutis@student.umass.edu
(17:29:40) dr ttol: ruth1783
(17:29:49) David Gucwa: I'm getting umich, cornell, yale and harvard
(17:30:06) dr ttol: ok
(17:30:07) dr ttol: get umass next
(17:30:40) David Gucwa: k
(17:31:00) dr ttol: we want to import the top i2hub schools first
(17:31:48) David Gucwa: ok
(17:31:54) David Gucwa: well I've used all the logins you gave me
(17:32:09) David Gucwa: so if you give me more logins I can start importing
(17:32:13) dr ttol: ok
(17:32:14) dr ttol: you mean
(17:32:17) David Gucwa: also I'm waiting for wellesley to start showing emails again